



National Environmental Scientific Computing Center User Guide

Version 8.0

<http://www.epa.gov/nesc/>

Contents

Quick Reference

This document is divided into the following sections:

1. Quick Reference
2. About the National Environmental Scientific Computing Center
3. Hardware at NESC²
4. Cray T3E-1200
5. IBM RS/6000 SP
6. Processes Common to Both HPC Systems
7. Customer Support Services

Please send any corrections, comments or suggestions concerning this document to help.nesc@epa.gov.

1.1 NESC² User Hotline

Monday-Friday, 8:00 a.m. - 5:00 p.m. (Eastern): **(919) 541-7862**; 1-800-334-2405

1.2 Internet Addresses (numeric & host names)

Cray T3E - Name: hickory.nesc.epa.gov; IP address: 134.67.65.45

IBM RS/6000 SP -Name: cypress00.nesc.epa.gov; IP address: 134.67.68.20

Name: cypress01.nesc.epa.gov; IP address: 134.67.68.21

Name: cypress02.nesc.epa.gov; IP address: 134.67.68.24

1.3 US EPA Manager Contact Information

The National Environmental Scientific Computing Center (NESC²) is a contractor-operated, EPA-managed facility designed to support EPA high performance computing programs. All requests for accounts and HPC support must be sent to the US EPA *National Platform Manager for High Performance Computing, Scientific Visualization, and Scientific Applications*:

John B. Smith
National Environmental Scientific Computing Center/NTSD
National Computer Center
U.S. EPA, Research Triangle Park, NC 27711

E-mail: smith.johnb@epa.gov;
Telephone: 919-541-1087

1.4 Support E-mail Addresses

System Administrator: sysadmin.nesc@epa.gov

Tape Librarian: tapelib.nesc@epa.gov

User Help: help.nesc@epa.gov

1.5 Online Documentation

This User Guide is available in Adobe Portable Document Format (PDF) on US EPA's Web site at

http://www.epa.gov/nesc/10_publications/Customer_Document/UserGuide.pdf

Additional user documentation is also available at

http://www.epa.gov/nesc/10_publications/Customer_Document/

About the National Environmental Scientific Computing Center (NESC²)

2.1 About this User Guide

The purpose of this guide is to provide enough information for a user to initially access and log into the National Environmental Scientific Computing Center's (NESC²) principal computing systems and to help users locate appropriate online sources of more detailed system and application information.

Information in this document is subject to change and users are encouraged to consult the most current revision. Access to this guide, and other online documentation, is described in later sections.

2.2 Mission of the NESC²

The mission of NESC² is to support the Agency mission by providing reliable, responsive, scalable, flexible, and cost-effective computing solutions for its high-end information processing, information management, and visualization requirements.

The business of NESC² is to provide Agency customers with high performance computing, scientific information management, and scientific visualization services, resources, and expertise, to support their progress towards key Agency goals.

2.3 How to Get an Account on an HPC System

NESC² accounts are available in two major forms: *Trial* (short-term) accounts, and *Project* (long-term) accounts. These types of accounts, and others, are available to EPA researchers, EPA contractors, and individuals affiliated with EPA-sponsored projects.

Applications for **project accounts** are solicited in early spring and are evaluated by the High Performance Computing Working Group (HPCWG) during the summer. Allocations of system resources to Project accounts are finalized and begin at the new fiscal year (October 1).

Please refer to the EPA NESC² Web page:

http://www.epa.gov/nesc/00_general/account.html

for a detailed description of how to apply for a Project account.

Inquiries regarding **trial accounts** may be directed to the US EPA manager. Refer to section 1.3 for contact information. To make an application for a *trial account* please send the following data about yourself and project to the US EPA manager via EPA, Internet, or U.S. mail:

1. Name
2. EPA 3-character login-ID (if available)
3. U. S. Mail address
4. EPA E-mail address (if available)
5. Internet E-mail address (if available)
6. Phone number
7. A brief description, in layman's language, of the application for which you need the supercomputer. Please limit this description to a one page narrative without mathematical formulas.
8. A cover letter from the Laboratory Director of the appropriate EPA facility.

About the National Environmental Scientific Computing Center (NESC²)

(Continued)

2.4 Accessing NESC² HPC Systems

Access to the NESC² is through TCP/IP using either the Internet or dedicated high-speed lines connected to EPA's internal network. To reach the IBM SP, enter:

```
telnet cypress00.nesc.epa.gov    Or
telnet cypress01.nesc.epa.gov    Or
telnet cypress02.nesc.epa.gov
```

To reach the T3E, enter:

```
telnet hickory.nesc.epa.gov
```

You are then prompted to log in. If you encounter any difficulties while logging in, contact the User Hotline at (919) 541-7862 or through the e-mail address:

help.nesc@epa.gov.

2.5 Finding Information on the HPC Systems

One goal at the NESC² is to have as much system and application documentation as possible available online to users:

- The message of the day (**motd**) displays important information when you log on to the Cray T3E or IBM RS/6000 SP.
- The **news** command is available on the T3E and IBM to display information. Note that at times there may not be any news items on one or both of these systems.
- UNIX **man** pages provide terse online technical documentation.
- [T3E only] The module display utility `<module help topic>` can be used to display specific information concerning use of the Cray T3E (hickory). Refer to section 4.2 for more information.
- The Cray Research Online Software Publications Library is now available on an internal EPA Web server at <http://linden.nesc.epa.gov/craydoc/frontpage.html>.

2.5.1 *news* Command

Current information on a number of topics is available through the **news** command. (This **news** is different from *newsgroups* on the Internet). The format of the news command is **news topic**.

The command **news items** provides a list of available topics. Entering **news** without a topic will send all unread news items to **stdout** (usually your display screen).

For hickory, news items are located in hickory in `/usr/news` and for cypress in `/var/news`. Important announcements, especially for cypress, are displayed to your screen when you log in.

2.5.2 *man* Command

The **man** command is used to access the set of **man[ual]** pages, which contain technical descriptions of UNIX commands, library calls, and other technical topics. The command,

```
man man
```

displays the *man page* for the **man** command. The **-k** option for **man** is recommended for performing keyword searches of the man pages. For example, to

About the National Environmental Scientific Computing Center (NESC²)

(Continued)

Hardware at NESC²

list all the UNICOS/mk performance-related commands, enter this command on hickory:

```
man -k performance | more
```

where entering | **more** displays the information one screen at a time. Press the spacebar to see the next screen. Press the "h" key to get help for using the **more** command. For example, you could read this guide with the command:

```
news guide | more
```

3.0 Hardware at NESC²

As visually described on the last page of this, NESC² brings together many different components in a sophisticated, high-bandwidth network. NESC² divides its HPC assets into three categories; it's assets are listed below within these three categories:

I. *Computational* Resources

- Cray T3E-1200E
- IBM RS/6000 SP
- 200+ Gflop/s Peak

II. *Data Management* Resources

- Sun E4500 File Server
- STK Tape Silos
- >90 TB Tape Capacity

III. *Networking* Resources

- Gigabit Ethernet - 1 Gb/s
- High Performance Parallel Interface (HiPPI) - 800 Mb/s
- Fiber Direct Data Interface (FDDI) - 100 Mb/s

3.1 Cray T3E

NESC²'s Cray T3E-1200E is equipped with 120 processing elements (PEs) and 256 MB of memory per PE. Each PE has a 600 MHz Alpha 21164 microprocessor with a peak speed of 1.2 billion floating point operations per second (1.2 Gigaflop/s). Each PE also has its own network router, connected in a dedicated 3-D torus topology with a peak speed of 650 MB/s per link in each direction. In its current configuration, the system has 112 PEs for application use, providing 28 GB of total memory and peak performance of 134 Gflop/s.

A dedicated CRAY ND-40 disk array provides fast access to 1.3 trillion bytes (1.3 TB) of online disk storage via a High Performance Parallel Interface (Hippi) connection for the T3E.

3.2 IBM RS/6000 SP

NESC²'s IBM RS/6000 SP high performance computer is equipped with three nodes, each with 16 375-MHz Power3-II processors and 16 GB of memory. Each processor has a peak speed of 1.5 Gflop/s, providing a total of 48 processors, 48 GB of memory, and a peak performance of 72 Gflop/s for application use. Cypress' nodes are connected via the high performance Colony switch, which can scale to 16 nodes as currently configured. The IBM SP provides EPA scientists with a

choice of parallel programming models, either a shared memory programming environment using up to 16 processors and 16 GB on a node, or a distributed memory programming environment using message-passing to communicate between processes on different nodes. Cypress also has 4 TB of online disk storage.

3.3 Sun E4500 Filer Server

To manage the large volume of data accessed and created by applications running on the IBM and Cray systems, NESC uses a Sun E4500 file server running the archival file system SAM-FS, which serves as the point of access for a tape archive containing over 78 TB of data.

Cray T3E-1200

4.1 Operating System (OS)

The operating system on the CRAY T3E-1200 is UNICOS/mk. It presents a single system image regardless of how many processors are installed on the system. All CRAY T3E systems have a mix of application (APP) processing elements (PEs), command (CMD) PEs, and operating system (OS) PEs. The APP PEs run the parallel jobs, while the CMD PEs run the single processor jobs and interactive commands and the OS PEs handle system tasks. To see a summary of the PE pools and what is currently running, enter the command `grmview`:

```
hickory 35% grmview
PE Map: 120 (0x78) PEs configured(normal scheduling)
      Ap. Size  Number Aps.  Abs.      <--- Lists --->
Type  PE  min  max running limit limit  Label svc uid gid acid
+ APP   0   2  112      1     1     2    -   -   -   -   -
  95 identical PEs skipped
+ APP 0x60   2  112      0     1     2    -   -   -   -   -
  15 identical PEs skipped
+ CMD 0x70   1   1      1 unlim unlim    -   -   -   -   -
+ CMD 0x71   1   1      1 unlim unlim    -   -   -   -   -
+ CMD 0x72   1   1      2 unlim unlim    -   -   -   -   -
+ CMD 0x73   1   1      1 unlim unlim    -   -   -   -   -
+ OS  0x74   0   0      0 unlim unlim    -   -   -   -   -
  2 identical PEs skipped
+ OS  0x77   0   0      0 unlim unlim    -   -   -   -   -
```

This display was taken from a day when 96 APP PEs were busy running user jobs. For more details about what is running on each processor and its x-y-z coordinate in the 3-D processor map, enter the command `grmview -l`.

If you want to see what version of UNICOS/mk we are running on the CRAY T3E, the command is

```
hickory 19% uname -a
sn6907 hickory 2.0.5.49 unicosmk CRAY T3E
```

4.2 User Environment

The recommended way to set up one's user environment on the CRAY T3E is using modules. The modules package contains commands for setting environment variables and search paths to simplify access to installed tools, libraries, and commands. Most accounts are created with module initialization commands already included in their startup files.

Cray T3E-1200*(continued)*

C shell users should find the following lines in their `.cshrc` file:

```
if (-f /opt/modules/modules/init/csh ) then
  source /opt/modules/modules/init/csh
  module load modules PrgEnv craytools mpt
endif
```

while K shell users should find the following in their `.profile` file:

```
if [ -f /opt/modules/modules/init/ksh ]; then
  . /opt/modules/modules/init/ksh
  module load modules PrgEnv craytools mpt
fi
```

It is important not to overwrite any of the path and environment variables settings after the module load command has been executed. If you want to add a search path, be sure to retain the previous settings. For example, if you are a C-shell user and wish to add `/usr/bin/x11` to your path, add the following line to your `.cshrc` file: `set path=($path /usr/bin/X11)`

You can manage your software environment through use of module commands. The available commands are

```
module list
    Lists all loaded modules
module avail
    List all available modulefiles in the current MODULEPATH
module display <modulefile>
    Display changes modulefile will make to the environment
module switch <oldmodulefile> <newmodulefile>
    Switch loaded oldmodulefile with newmodulefile
module load <modulefile>
    Load modulefile from the shell environment
module rm <modulefile>
    Remove modulefile from the shell environment
module help <modulefile>
    Prints the module specific help information for the modulefile.
```

For example, the command to use the tools, libraries, and compilers from Programming Environment 3.5, instead of the default PrgEnv environment, is as follows:

```
module switch PrgEnv PrgEnv.35
```

4.3 Examining Your Job and Account Limits

Users may examine their per process and job account limits by using the `udbsee` command. See news item *user_limits*, and the `udbsee` and `udbgen` man pages for details. In particular the man page for `udbgen` describes all possible fields for user resource limits.

To allow users to view project accounting and allocation information, the NESC² provides the `ru` command. `ru` (resource utilization) is a useful command for viewing your CPU usage FY-to-date. It also shows the usage of your group against your group's allocation. Allocations are assigned annually by the High Performance Computing Working Group, following peer review. One-fourth of the yearly allocation is allotted at the beginning of each quarter. If your group exceeds its

Cray T3E-1200*(continued)*

allocation in a quarter, your jobs will still run if resources are available, but at a lower priority. See the man page for `ru` for more information.

4.4 User File Systems

The active user file systems on the CRAY T3E are

- /home** user home directories. `/home` is intended for small personal files such as startup and configuration files and scripts. Each user has approximately 150 MB of space.
- /work** group work directories. `/work` is intended for work space and active copies of programs and data files. Files placed in work remain online and do not migrate, so if your quota is used up you need to work with other members of your group to free up space.
- /asm** archival file system. `/asm` is actually two file systems, `/asm1` and `/asm2`, running on the Sun file server and NFS-mounted to both the CRAY T3E and the IBM SP. Files copied to `/asm` are initially held in an online "disk cache" on the Sun server but are sent to tape upon arrival.

Both `/home` and `/work` are managed by quotas. Each user has an individual quota for `/home`, and shared a quota for `/work` with other members of his or her group. To view your current quota limits for online file space, enter the command `quota`.

The archival file systems previously managed by the CRAY C90 (sequoia) under the Data Migration Facility (DMF) are still accessible on the CRAY T3E, but have been modified to be read-only. Their files are in the process of being transferred to the `/asm` file system. The old sequoia file systems are:

- /archive** file system previously shared between the CRAY C90 and the CRAY T3E, created in 1999
- /sequoia/home** old `/home` from sequoia
- /sequoia/work** old `/work` from sequoia

In addition, two temporary file systems are available for extra work space:

- /tmp** System temp file system, purged automatically as it fills up
- /ptmp** "Permanent" temp space, files are not purged but access is limited to projects needing a larger work area.

4.5 Programming Environment

The tools for compiling and linking programs on the CRAY T3E are

- f90** Fortran 90 compiler, compatible with the Fortran 95 standard, with extensions for co-array Fortran
- cc** ANSI standard C compiler
- CC** C++ compiler, supporting the ISO/ANSI Draft Working Paper for C++
- cld** Cray loader, usually invoked automatically by **f90**, **cc**, or **CC**

Cray T3E-1200*(continued)*

You can see which version of the compiler you are using with the `-v` option, for example:

```
hickory 58% f90 -V; cc -V; CC -V
Cray Fortran: Version 3.5.0.1 Thu Aug 15, 2002 14:17:55
Cray Standard C Version 6.5.0.1 (u138c34093p42027m32034a15) 08/15/02
14:17:56
Cray C++ Version 3.5.0.1 (u138c34093p42027m32034a15) 08/15/02 14:17:57
```

The Application Programming Interface (API) for all Cray systems is 64-bit, meaning that all pointers are 64 bits long, but the effective memory limit is approximately 240 MB, the maximum memory available on a CRAY T3E processor after allowing for the operating system microkernel running on each node. To access more memory, you need to use more nodes (a CRAY T3E node consists of one processor, its memory, and its network routing hardware).

Most programs targeted for the CRAY T3E use a distributed memory programming environment. Support is provided for:

- MPI** (Message Passing Interface): Cray's Message Passing Toolkit provides support for MPI version 1.2
- PVM** (Parallel Virtual Machine): including both the PVM message passing library and the PVM console.
- Shmem**: CRAY's native Shared Memory Library, which provides support for low-level one-sided communication routines
- CAF** (Co-Array Fortran): an emerging standard for expressing program parallelism without explicit message passing. Recognition of Co-Array Fortran syntax is provided with the `-z` option to the `f90` command.

The MPI, PVM, and Shmem libraries are linked automatically with user programs by default. The number of processors to be used at run time can be specified at compile time via the `-x <npes>` option to the compiler, or it can be specified at run time using options to `mpprun` or `mpirun`:

```
mpprun -n 16 a.outOr
mpirun -np 16 a.out
```

4.6 Data Format Conversion

The CRAY T3E provides automatic data conversion on input or output as an option to the Cray-specific `assign` command. The `assign` command associates file characteristics with a Fortran unit number for use during the library open processing. It is issued one or more times before the executable, as follows:

```
assign <options> <assign_object>
assign <options> <assign_object>
a.out
```

Options are assigned whenever a file is opened, whether through an explicit Fortran `OPEN` command, an implicit open via some Fortran I/O statement, or through a library routine such as `ffopen` or `AQOPEN`. The `<assign_object>` argument takes one of the following forms:

| Format | Example | Attribute association |
|---------------|---------|---------------------------------------|
| g:io_type | g:su | Sequential unformatted open request |
| u:unit_number | u:9 | Fortran unit 9 |
| p:pattern | p:file% | File names matching the pattern file% |
| f:file_name | f:file1 | File name file1 |
| file_name | myfile | File name myfile |

Cray T3E-1200*(continued)*

Some of the more commonly used <options> are

```
-a <filename>  associate a file name (usually, with a unit number)
-s <filetype>  associate a file type, such as cos or unblocked
-f 77          use Fortran 77 style namelist (a Cray extension)
-N <type>      convert files from the indicated type on input,
               or to the indicated type on output.  The two most
               commonly seen conversions are -N cray and -N ieee.
```

Assignments are stored in the file `$TMPDIR/.assign` and remain active until you remove them or log out. An assignment of options to the same object as a previous assign will replace the previous assignment unless the `-i` option is specified to indicate an incremental assign. Two useful commands for managing the `.assign` file are

```
assign -R      Remove all current assignments
assign -V      View all current assignments
```

Many projects at NESC avoid data conversions by writing their files in a common data format such as netCDF (<http://www.unidata.ucar.edu/packages/netcdf/>). The netCDF library and tools are installed on both the CRAY T3E (via module `netcdf`) and the IBM SP (in `/usr/local/lib`).

4.7 Batch Queueing System

Batch job processing on the CRAY T3E-1200 is provided by the Network Queueing Environment (NQE). This system queues jobs until resources are available and includes capabilities for scheduling and prioritizing jobs and managing the workload. It also includes an automatic checkpoint/restart feature that allows jobs to continue processing near where they left off in the event of planned or unplanned down time.

The main NQE commands are

```
qsub <options> <script> : submit a job to the queues
qlimit                  : summarize all qsub options
qstat                   : list the available queues
qstat -f <queue_name>   : view the queue limits
qstat -a                : track my queued jobs
qdel <requestid>        : delete a job from the queues
```

Most options to the `qsub` command can be specified in the job script itself, so submitting a job can be as simple as `qsub <script>`. Here is a sample `qsub` script requesting 12 processors and one hour of run time:

```
#QSUB -eo
#QSUB -l mpp_p=12
#QSUB -l mpp_t=1:00:00
#QSUB -l p_mpp_t=1:00:00
#
cd $HOME/linpack
mpirun -np 12 linpack
```

Currently there is one queue for each project on the CRAY T3E, as well as queues for long single processor jobs and special projects. Contact the NESC staff for assistance in choosing the right queue for your work.

IBM SP

5.1 Operating System

The operating system on the IBM SP is AIX, currently version 4.3.3. Each node of the IBM SP appears as an independent system, connected by a high-speed switch. The nodes are named cypress00, cypress01, and cypress02, and each node has its own IP address. Some file systems are shared across nodes, such as `/gpfs/home` and `/work`, while other file systems are local to the node, such as `/tmp` and `/ptmp`. For more details on the file system organization, see section 5.6.

If you want to see what version of AIX installed components we are running on the IBM SP, the command is `lslpp`, for example:

```
cypress01 241% lslpp -L | grep "bos\.mp"

bos.mp                                4.3.3.78      C      Base Operating System
```

5.2 User Environment

Modules are not available on the IBM SP. Default settings for the `PATH` and `MANPATH` environment variables for all users are found in the system file `/etc/environment`. If you want to add to these paths, be sure not to overwrite the system settings. For example, to add NCAR Graphics and netCDF to your search path every time you log in, you could put the following lines in your `.cshrc` file:

```
setenv NCARG_ROOT /usr/local/ncarg
setenv NETCDF_PATH /usr/local/lib/netcdf-3.4
setenv PATH "${PATH}:${NCARG_ROOT}/bin:${NETCDF_PATH}/bin"

setenv MANPATH "${MANPATH}:${NCARG_ROOT}/man:${NETCDF_PATH}/man
```

5.3 User File Systems

On the IBM SP, most system file systems are local to the node but most user file systems use IBM's General Parallel File System (gpfs) and are shared by all the nodes. The user file systems on the IBM SP are

- `/gpfs/home` GPFS file system for user home directories. All users share approximately 70 GB. Quotas are not enforced.
- `/work` GPFS file system for user work directories. All users share approximately 2.3 TB—quotas are not enforced.
- `/ptmp` local (non-GPFS) file system, available on request for projects that need a larger work area. Transfer rates are faster than to `/work`.
- `/asm` archival file system, runs on the Sun file server

Users accustomed to the Cray environment may be tempted to use the `/tmp` file system, but this is a small (6 GB) non-shared space needed by the system and should not be used as a work area.

IBM SP

(continued)

5.4 Programming Environment

The IBM Fortran and C compilers are called `xlf` and `xlc`, respectively, but there are many different ways to invoke them which include different combinations of compiler and loader options. They are

| | |
|----------------------------|---|
| <code>xlf</code> | Fortran compiler with fixed form default, as in Fortran 77 |
| <code>f77</code> | Alias for <code>xlf</code> |
| <code>xlf90</code> | Fortran 90 compiler (free form default) |
| <code>xlf95</code> | Fortran 95 compiler |
| <code>xlf_r</code> | <code>xlf</code> with links to thread-safe components |
| <code>xlf90_r</code> | <code>xlf90</code> with links to thread-safe components |
| <code>xlf95_r</code> | <code>xlf95</code> with links to thread-safe components |
| <code>mpxlf</code> | <code>xlf</code> behavior when using MPI |
| <code>mpxlf90</code> | <code>xlf90</code> behavior when using MPI |
| <code>mpxlf95</code> | <code>xlf95</code> behavior when using MPI |
| <code>mpxlf_r</code> | <code>xlf_r</code> behavior when mixing MPI and threaded programming |
| <code>mpxlf90_r</code> | <code>xlf90_r</code> behavior when mixing MPI and threaded programming |
| <code>mpxlf95_r</code> | <code>xlf95_r</code> behavior when mixing MPI and threaded programming |
| <code>mpxlf_chkpt</code> | <code>mpxlf</code> with parallel checkpoint/restart |
| <code>mpxlf90_chkpt</code> | <code>mpxlf90</code> with parallel checkpoint/restart |
| <code>mpxlf95_chkpt</code> | <code>mpxlf95</code> with parallel checkpoint/restart |
| <code>xlc</code> | ANSI standard C compiler |
| <code>cc</code> | Alias for <code>xlc</code> |
| <code>mpcc</code> | C compiler with links to MPI |
| <code>mpcc_r</code> | C compiler with links to threaded MPI and the Low-level Applications Program Interface (LAPI) |
| <code>mpcc_chkpt</code> | <code>mpcc</code> with parallel checkpoint/restart |

You can see which version of the compiler is installed by looking at the output of the `lslpp` command, for example:

```

cypress01 100% lslpp -L | grep "xlf rte"
xlf rte                7.1.0.2    C      XL Fortran Runtime
xlf rte.aix43          7.1.0.2    C      XL Fortran Runtime Environment
xlf rte.msg.en_US      7.1.0.0    C      XL Fortran Runtime Messages -
cypress01 108% lslpp -L | grep "vacpp\ .cmp\ .rte"
vacpp.cmp.rte          5.0.2.0    C      VisualAge C++ Compiler

```

There are two Application Programming Interfaces (APIs) supported on the IBM SP. The default is a 32-bit API, in which all pointers are 32-bits long and the addressable memory space is limited to 2 GB. This is sufficient for most MPI programs because we only have a total of 16 GB of memory available for the 16 processors on a node. The 64-bit API is also available through the compiler option `-q64`; this option changes the size of pointers to 64 bits. The size of other variables, such as floating-point or integer variables, is not changed by the addition of the `-q64` flag.

The IBM SP supports both a shared memory programming model using the processors and memory space of a single node, and a distributed memory programming model, using 1 or more nodes. The `-qsmp` option to `xlf` or `xlc` enables compiler recognition of shared memory parallel (SMP) directives, including OpenMP directives. The `mp---` compilers provide support for MPI programs. The IBM Parallel Environment for AIX supports MPI version 1.2 and some portions of the MPI-2 standard.

Execution of parallel programs is initiated by the IBM-specific `poe` command, which stands for *Parallel Operating Environment*. When running interactively, `poe` looks

IBM SP*(continued)*

for a file containing a list of host names, one per process, specifying the node on which to run each process of the parallel job. For example, an interactive parallel invocation of the “hello, world” program might look like

```
cypress01 113% poe hello_mpi -hostfile hostfile -procs 8
```

with the file “hostfile” containing at least 8 lines such as

```
cypress00
cypress00
cypress00
cypress00
cypress00
cypress00
cypress00
cypress00
cypress00
```

Alternatively, one can set the environment variable `MP_HOSTFILE` equal to the name of an existing host file and leave off the `-hostfile` option.

When `poe` is invoked in a batch environment, a system-generated host file is used.

5.5 Data Format Conversion

The IBM SP does not have an equivalent of the `assign` command (as on the T3E) and so does not support automatic data conversion. Instead, any “foreign” data conversions required on input must be done by reading the raw data into a buffer and converting it to the desired format by calling library routines. A common usage of these library routines is converting datasets written in binary format on a CRAY vector machine, such as the CRAY C90 that was in use at NESC² from 1994-2001, to IEEE format, which is native on the IBM SP. The NCARU library (<http://www.scd.ucar.edu/docs/conversion.tools/ncaru.html>) has been installed on the IBM SP for this purpose.

5.6 Batch Queueing System - LoadLeveler

The batch queueing system on the IBM SP is called *LoadLeveler*. It is used for scheduling jobs and dividing the workload among all the nodes in the system. There is no automatic checkpointing for most jobs but some variants of the compiler claim to support this functionality. Please contact the NESC staff for assistance in implementing checkpointing for your jobs.

The main LoadLeveler commands are

```
llsubmit <script>      : submit a job to the queues
llstatus               : show status of the nodes
llclass               : list the job classes (queues)
llclass -l <queue_name> : view the queue limits
llq                   : show the queued and running jobs
llq -u <username>      : only view one user's queued jobs
llcancel <requestid>   : cancel a job in the queues
```

Here is a sample LoadLeveler job script:

```
#!/bin/csh
#
# @ error    = /dev/null
# @ output   = cypress.16.${jobid}
# @ notification = never
```

IBM SP*(continued)*

```
# @ environment = MP_EUILIB=ip
# @ wall_clock_limit = 1:00:00
# @ job_type = parallel
# @ node = 2
# @ total_tasks = 16
# @ network.mpi = css0,shared,US
# @ class = medium
# @ node_usage = shared
# @ queue
```

```
poe linpack
```

The environment variable setting of `MP_EUILIB=ip` is the default and is just included for illustration.

Processes Common to Both HPC Systems

6. Processes Common to Both HPC Systems

6.1 How to Change Your Password

When you first log in to a new account on a NESC² system, you must immediately select a password for your account. Establishing a new password differs on the IBM SP and Cray T3E. Refer to the corresponding section below for the correct procedure.

After your initial password selection, you will be required to change your password every 90 days. You will use the `passwd` command to do this.

Note: Separate accounts are required for hickory and cypress. An account on cypress00, cypress01, or cypress02 provides access to all three of the SP's nodes (and the operator workstation, "juniper").

Note: EPA policy prohibits the sharing of accounts/passwords under any circumstances.

6.1.1 Changing Your Password on the Cray T3E

You will be asked to retype the old (original) password, and are then presented with a random pseudo-pronounceable password that has been generated for you. If this password is not to your liking, you may press the carriage return key and have the password generator present a different choice. However, you must use a system-generated password, not one that you make up. ***Be sure to note the new password immediately since you will not be able to log in again without it.*** After resetting your password for the first time, you will be disconnected and must re-connect and log back in using the new password.

6.1.2 Changing Your Password on the IBM SP

Changing your password by logging into one of the three SP nodes effects the change only for about 10-15 minutes. After that, the system resets your password to the old one. To change your password in a way that the SP not only retains the change but also propagates that change to all three nodes, you must log in to the operator's workstation for the IBM SP, which is called *juniper*. You can only reach *juniper* from one of the cypress nodes. Use your old password to log in, even if you have just changed it on one of the nodes. On *juniper*, run the `passwd` command and enter your old password as the old password. Your new password can be anything

Processes Common to Both HPC Systems

(Continued)

you choose; it is not generated for you as on the CRAY T3E. Once the new password propagates to the nodes, you will have the same password on cypress00, cypress01, cypress02, and juniper.

2.6 Forgot Your Password? Here's How to Get it Reset

One of the most common user requests is to have their password reset because they've forgotten it. Of course, the obvious solution is to always remember your password. However, we all forget our password sooner or later. Here's how to have it reset. By the way, NESC² staff cannot reset a user's password simply by their asking due to established security policies.

If you forget your password, you must initiate the following six step process (1) request a *password reset* from your High Performance Computing Working Group (HPCWG) representative. (2) Your HPCWG representative will convey your request to the EPA Task Order Project Officer (identified in section 1.3), who will issue technical direction (3) to the NESC staff to reset your password. NESC staff will issue the new password to the TOPO (4), who will give it to your HPCWG rep (5), who will give it to you (6). And you're back in business.

6.2 Using the /asm File System

6.2.1 Putting Files on /asm

Although **/asm** looks just like a regular UNIX file system and supports UNIX commands to access files stored there, some functions are not advised for performance reasons. In particular, writing to a file on **/asm** is very slow—less than 1 MB/second. Instead, users should create all their output files on locally managed file systems and copy the complete file to **/asm**. Options for storing a temporary file locally include the **/tmp** and **/work** file systems on the CRAY T3E and the **/work** file system on the IBM SP. NESC staff can work with individual users on setting up additional scratch areas if these file systems prove inadequate for their applications.

After you've created a file on hickory or cypress, you should use the NESC command **archput** to copy it to **/asm**. The simplest usage is

```
archput <filename>
```

This command copies **<filename>** to a corresponding directory on **/asm**. The **archput** command tries to replicate the current directory structure for **<filename>** on **/asm** by copying files from **/home** or **/work** to **/asm** and files from **/sequoia/home** or **/sequoia/work** to **/asm/sequoia/home** or **/asm/sequoia/work**. For example,

```
/home/gah/<filename>          is copied to /asm/gah/<filename>
/work/gah/sub/<filename>      is copied to /asm/gah/sub/<filename>
/sequoia/work/gah/<filename> is copied to /asm/sequoia/work/gah/<filename>
```

If the file already exists on **/asm**, it is overwritten. You can also specify an alternate directory to **archput** using the **-a** option. The command

```
archput -a newdir <filename>
```

will copy **<filename>** to **/asm/gah/newdir** for user "gah" if the file comes from **/home** or **/work**, or to **/asm/sequoia/work/gah/newdir** if the file comes from **/sequoia/work**. If "newdir" is a fully qualified path name, the file will always be copied there regardless of where it currently resides.

Processes Common to Both HPC Systems

(Continued)

The `archput` command can also take a list of files from standard input for use in a UNIX pipe. For example, you could copy all the files in the directory `mydir`, including any subdirectories, with the commands

```
cd /work/gah/mydir
find . -type f -print | archput
```

Note that the `archput` command copies files to `/asm`, but it doesn't delete the original file. You should verify that the sizes match, and perhaps the checksums as well, before deleting the original file.

6.2.2 A Note About FTP (File Transfer Protocol)

Some users have reported problems when using `ftp` to transfer files from outside the NESC domain to cypress or hickory if the destination file is on `/asm`. Don't do this. If you need to transfer files via `ftp` directly to `/asm`, you should request a password for the Sun file server (sweetgum) from your SWG representative.

6.2.3 Reading Files from /asm

Former sequoia users are well-acquainted with the Data Migration Facility (DMF) command `dmget` to recall a migrated file from tape. The equivalent command for the `/asm` file system is called `archget`. Like `dmget`, it causes one or more files to be staged from the archive media to online disks for faster access. The syntax is

```
archget <filelist>
```

where `<filelist>` Specifies a list of files to recall from `/asm`. If you don't specify `<filelist>`, the command reads from standard input.

Read performance from `/asm` to the IBM SP is competitive to that of local disks, so it is best to leave the file where it is for reading. However, read performance from `/asm` to the CRAY T3E is much slower than reading from local disks, so it is often beneficial to copy the file locally. The `archget` command has a `-c` option to specify a local directory for the copy:

```
archget -c <cdir> <filelist>
```

If `<cdir>` is not a full path name, the files are copied to `/work/gah/<cdir>` for user `gah`.

6.2.4 Transition from DMF to ASM

Some users have noticed that they have other directories on `/asm` under

```
/asm/sequoia/home
/asm/sequoia/work
```

These directories are a work in progress to transfer the DMF files that formerly resided on sequoia to `/asm`. The `/asm/sequoia/work` directory is not complete and users needing access to their sequoia files should access them from `/sequoia/work` on hickory.

6.3 Backup Procedures

All user data in `/home` on the CRAY T3E are backed up daily. Data in `/work` and `/archive` on the T3E are backed up weekly. The daily backups are incremental, recording only those files modified or created since the last backup. Weekly backups are full backups, including a complete copy of each file, except that only

Processes Common to Both HPC Systems

(Continued)

the i-nodes (the file metadata) are backed up for migrated files from `/archive`. The last full backup of each month is sent off site and kept for six months.

On the IBM SP, a full backup of the nearly 3 TB of online disk space is not feasible at the present time. Instead, a single copy of the online data is kept on `/asm` and updated weekly. All users are encouraged to make their own additional backup copies of important files by copying them to `/asm` as needed.

There is only one copy of all files sent to tape on `/archive` or `/asm`. Only the i-nodes are backed up on the Sun file server.

6.4 Debuggers

The source level debugger available on both the CRAY T3E and the IBM SP is called `totalview`. The `totalview` debugger is a Cray product for the CRAY T3E but is provided by Etnus for the IBM SP, so the graphical user interface looks slightly different on the two systems.

To use the debugger, you must compile your program with the `-g` option to both the CRAY and IBM compilers. To run the debugger interactively, enter `totalview` on the command line and choose "Load new program" or "New program" from the File menu. You can also specify the core file or program to run from the `totalview` command line. Online help is available from the TotalView window.

6.5 E-mail on the Cray T3E

In addition to large scale scientific computing support, NESCS² also offers each user access to electronic mail on the Cray T3E. NESCS² staff strongly recommends that users receive and process their e-mail on their own PC or workstation—where much better mail tools exist, rather than on the Cray. Users should specify their normal e-mail address in a ".forward" file in their home directory on the Cray, e.g., `whitman.christine@epa.gov`.

For those users who must process their e-mail on the Cray, the `mailx_tutorial` (`news mailx_tutorial`) describes the preferred e-mail program for UNICOS/mk: `mailx`.

6.6 How to Get Your Data on the System

The following sections describe the methods available for moving your data.

6.6.1 Tapes

Normal magnetic tape support for NESCS² is provided by the StorageTek 4400 silos. NESCS² is currently in the process of converting its archival storage media to STK 9840 and STK 9940 tapes.

Internally, NESCS² can provide limited support for 1/4 inch cartridge tapes, 4mm DAT tape, Exabyte 8mm helical scan tapes, and DDs-3 tapes (on juniper). Contact the user hotline for information on possible access to these tape formats.

6.6.2 Getting Tapes to the NESCS²

Send your tapes to:

Tape Librarian
National Environmental Scientific Computing Center
National Computing Center

Processes Common to Both HPC Systems

(Continued)

U.S. Environmental Protection Agency
Research Triangle Park, NC 27711

Users are responsible for reading, writing, and converting their own data from tapes.

6.6.3 Data Format Conversion

Many projects at NESC avoid data conversions by writing their files in a common data format such as netCDF (<http://www.unidata.ucar.edu/packages/netcdf/>). The netCDF library and tools are installed on both the CRAY T3E (via module `netcdf`) and the IBM SP (in `/usr/local/lib`).

6.6.4 Electronic Transfers

Most users will transfer their program source code and much of their data electronically. However, very large data sets should be transported with magnetic tape. The primary means for electronic transfers is the UNIX File Transfer Protocol (FTP).

Note: Remote NSF mounts are not supported by NESC².

In addition to standard FTP, UNICOS/mk offers additional capability to users through **ftua**, the *file transfer user agent*. The **ftua** utility allows a user to batch and queue file transfers, and provides retry facilities. See news item *ftp*, the man page for **ftua** and the *UNICOS TCP/IP User's Guide* for further information.

Large file transfers may also benefit from file compression utilities. See news item *file_compression*, and the man page for **pack** and **compress** for further information. The **gzip** utility is also available on the T3E and SP. For quick help with this compression command, enter **gzip --help**.

See the *Application Programmer's I/O Guide (SG-2168)* for more information on file and data conversion utilities.

Customer Support Services

7. Customer Support Services

7.1 Hours of Operation

Operations support is available 24 hours per day, seven days a week. The normally scheduled time for system maintenance are as follows: Mondays from 5:00 - 8:00 a.m. (Eastern Time).

Daily notices of changes to system availability will appear at user login in the *Message of the Day*.

Important: It is a good idea to read the message of the day every time you log in. See the man page **motd** for more information.

7.2 Who to Call for Help

The primary point of contact for users is the User Hotline. The Hotline telephone number is **919-541-7862**. Help is also available through the E-mail address: **help.nesc@epa.gov**.

Customer Support Services

(continued)

The Hotline is staffed from 8:00 a.m. to 5:00 p.m. Eastern Time, Monday-Friday. Users should direct inquiries, and problems related to communications, operations, hardware and software through the Hotline.

When the Hotline is not in service, operations problems may be directed to the NESC² operators' console: 800-334-2405. E-mail may be sent at any time to the e-mail addresses listed at the start of this document, under the heading *Quick Reference*.

7.3 Code Optimization

Because NESC²'s computational resources are limited, USA EPA requires that codes running on the Cray T3E or IBM SP be optimized for the system on which they run. Current account holders may contact Ed Anderson (contractor) at 919.541.0299 or anderson.edward@epa.gov for more information about NESC²'s code optimization services. Non-account holders interested in this service should contact the EPA manager identified in section 1.3.

7.4 Code Porting

NESC² staff is also available to assist customers in porting code from one programming language to another or from one computing platform to another. Current account holders may contact Ed Anderson (contractor) at 919.541.0299 or anderson.edward@epa.gov for more information about NESC²'s code porting services. Non-account holders interested in this service should contact the EPA manager identified in section 1.3.

7.5 Computational Chemistry

NESC² staff provide computational chemistry expertise. Current account holders may contact Dr. Charles Foley (contractor) at 919.541.1184 or foley.charles@epa.gov for more information about NESC²'s computational chemistry services. Non-account holders interested in this service should contact the EPA manager identified in section 1.3.

7.6 Computational Fluid Dynamics

NESC² also has on staff specialists in computational fluid dynamics (CFD). These staff are familiar with applying Navier-Stokes equations, available in commercial-off-the-shelf (COTS) applications, such as Fluent, to scientific problems such as fluid through dense urban areas and the human respiratory system. NESC² staff can also create custom CFD applications based on available funding. Current account holders may contact Dr. Matt Freeman (contractor) at 919.541.4293 or freeman.matt@epa.gov for more information about NESC²'s CFD services. Non-account holders interested in this service should contact the EPA manager identified in section 1.3.

7.7 Scientific Visualization

A key part of the National Environmental Scientific Computing Center is its Scientific Visualization Center (SVC). SVC offers a wide variety of services and these are described on its Web site at <http://www.epa.gov/vislab/>. Please contact the EPA manager identified in section 1.3 if you are interested in NESC²'s visualization services.

**Customer
Support
Services**

(continued)

7.8 NESC² User Training

NESC² staff can provide training for users. For information about available training, see the news item *NESC_training*. For any requests concerning training, EPA users should contact the NESC² US EPA manager. Refer to section 1.3 for contact information.

National Environmental Scientific Computing Center

